SPECTRAL SUBTRACTIVE TYPE SPEECH ENHANCEMENT METHODS

Ekaterina VERTELETSKAYA[1], *Boris ŠIMÁK*[2]

[1] Dept. of Telecommunication Engineering, Czech Technical University in Prague, Technická 2, Czech Republic
[2] Dept. of Telecommunication Engineering, Czech Technical University in Prague, Technická 2, Czech Republic

verteeka@fel.cvut.cz, simak@email.cz

Abstract *In this paper spectral subtractive method and some of its modification are compared. Performance of spectral subtraction, its limitations, artifacts introduced by it, and spectral subtraction modifications for eliminating these artifacts are discussed in the paper in details. The algorithms are compared based on SNR improvement introduced by them. Spectrograms of speech enhanced by the algorithms, which show the algorithms performance and degree of speech distortion, are also presented.*

Keywords

Noise reduction, Spectral subtraction, Musical noise

1. Introduction

The speech processing systems used to communicate or store speech are usually designed for a noise free environment but in a real-world environment, the presence of background interference in the form of additive background and channel noise drastically degrades the performance of these systems, causing inaccurate information exchange and listener's fatigue. Speech enhancement algorithms attempt to improve the performance of communication systems when their input or output signals are corrupted by noise. The main objective of speech enhancement or noise reduction is to improve one or more perceptual aspects of speech, such as the speech quality or intelligibility. It is usually difficult to reduce noise without distorting speech and thus, the performance of speech enhancement systems is limited by the tradeoff between speech distortion and noise reduction. The complexity and ease of implementation of any proposed scheme is another important criterion especially since the majority of the speech enhancement and noise reduction algorithms find applications in real-time portable systems like cellular phones, hearing aids, hands free kits etc. The numerous of speech enhancement techniques have been developed based on short-time spectral attenuation, speech modeling, wavelet transformation, and etc. [1] The spectral subtraction method has been one of the most well-known techniques for noise reduction. Due to its minimal complexity and relatively ease in implementation, it has been in the spotlight over the past years.

2. Basic principle of spectral subtraction

Spectral subtraction is build upon the assumption that the noise signal and the speech signal are uncorrelated signals added together to form the noisy speech signal [2]. The principle of the spectral subtraction method is based on estimating clean speech power spectrum by subtracting the noise power spectrum from the speech power spectrum that includes noise.We assume to have a speech signal *x(n)* corrupted by an additive noise *d(n)*. Then the received noisy signal *y(n)* is described by

$$y(n) = x(n) + d(n) \tag{1}$$

In the frequency domain, with their respective Fourier transforms, the power spectrum of the noisy signal can be represented as:

$$|Y(\omega)|^2 = |X(\omega)|^2 + |D(\omega)|^2 + X(\omega)\cdot D(\omega)^* + X(\omega)^*\cdot D(\omega) \tag{2}$$

,where $Y(\omega), X(\omega), D(\omega)$ are DFT magnitudes of *y(n), x(n), d(n)* respectively, $D(\omega)^*$ and $X(\omega)^*$ represent the complex conjugates of $D(\omega)$ and $X(\omega)$ respectively. If we assume that *d*(n) is uncorrelated with *x*(n) , then the terms $X(\omega)D(\omega)^*$ and $X(\omega)^*D(\omega)$ are reduced to zero. Power spectrum of the noise speech $D(\omega)$ cannot be obtained directly, but can be estimated during speech pauses (when *y(n)=d(n)*). The algorithm for separating conversational speech signal to speech and silence regions is called the voice activity detector (VAD). The estimation of noise signal power spectrum can be denoted by $|\hat{D}(\omega)|^2$. Thus from the above based assumptions, the estimate of clean speech can be given as (3):

$$|\hat{X}(\omega)|^2 = |Y(\omega)|^2 - |\hat{D}(\omega)|^2 \tag{3}$$

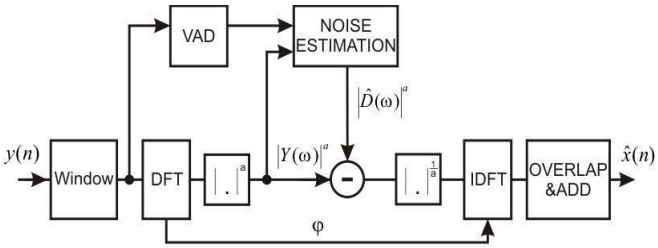

Fig. 1. General representation of spectral subtraction.

Alternatively a more general form is given by generalizing the exponent from 2 to *a*

$$\left|\hat{X}(\omega)\right|^a = \left|Y(\omega)\right|^a - \left|\hat{D}(\omega)\right|^a \qquad (4)$$

,where the power spectrum is exchanged for a general form of spectral density. Once the estimate of the clean speech is obtained in the spectral domain with the (4) the enhanced speech signal is obtained by inverse DFT transformation of $\hat{X}(\omega)$. Since the human ear is not sensitive to phase errors of the speech, the noisy speech phase can be used as an approximation to the clean speech phase, for reconstruction enhanced speech from its spectrum. Thus a general form of the estimated speech in frequency domain can be written as:

$$\hat{s}(n) = IDFT[(\left|Y(\omega)\right|^a - \left|\hat{D}(\omega)\right|^a)^{1/a}] \qquad (5)$$

Fig. 1 shows a block diagram of the spectral subtraction method. The processing, is carried out on a short-time basis (frame-by-frame), therefore, a time-limited window should be applied to input noisy speech signal at the beginning of the algorithm, and overlap add at the end is done to reconstruct the speech estimate in the time domain.

3. Noise estimation and speech/silence detection

A practical speech enhancement system consists of two major components, the estimation of noise power spectrum, and the estimation of speech. Therefore, a critical component of any frequency domain enhancement algorithm is the estimation of the noise power spectrum. In single channel noise reduction/speech enhancement systems, most algorithms require an estimation of average noise spectrum, and since a secondary channel is not available this estimation of the noise spectrum is usually performed during speech pauses. This requires a reliable speech/silence detector. The speech/silence detection

Scheme can be a determining factor for the performance of the whole system of noise reduction based on spectral subtraction. The speech/silence detection is necessary to determine frames of speech pauses or noise only frames, to allow an update of the noise estimate. If the

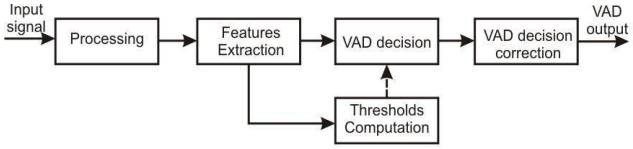

Fig. 2. Block diagram of a basic VAD design.

speech/silence decision is not correct then speech echoes and residual noise tends to be present in the enhanced speech. Typically, in recognizing the speech and noise segments of a speech signal, its energy level [3], pitch, zero crossing rate, statistical and spectral properties are used. The basic principle of a speech/silence detector is that it extracts measured features or quantities from the input signal and then compares these values with thresholds usually extracted from noise-only periods. Voice activity (VAD=1) is declared if the measured values exceed the thresholds. Otherwise, no speech activity or noise, silence (VAD=0) is present. Voice activity detector (VAD) tends to follow a common paradigm comprising a pre-processing stage, a feature-extraction stage, a threshold comparison stage, and an output-decision stage. A general block diagram of a VAD design is shown in Fig. 2.

4. Limitation of spectral subtraction

Noise spectrum estimate is obtained from the non-active regions of noisy speech. This assumption is valid for the case of stationary noise in which the noise spectrum does not vary much over time. Traditional VADs track the noise only frames of the noisy speech to update the noise estimate. But the update of noise estimate in those methods is limited to speech absent frames. This is not enough for the case of non-stationary noise in which the power spectrum of noise varies even during speech activity.

Spectral subtraction performance is limited by the accuracy of noise estimation, which additionally is limited by the performance of speech/pause detectors [4].VAD performance degrades significantly at lower SNR. However, the main problem with spectral subtraction is the processing distortions caused by random variations of the noise spectrum. Irrespective of the methods used for estimating the noise statistics, the true short spectrum of the noise will always have a finite variance. Thus the noise estimate will always be over or under the estimate of the true noise level. Therefore, wherever the noisy signal level is near the level of the estimated noise spectrum, spectral subtraction (4) results in some randomly located negative values for the estimated clean speech magnitude. To remove the negative components half-wave rectification (setting the negative portions to zero), or full wave rectification (absolute value) are used. The non-linear mapping of the negative, or small valued spectral estimates, results in the estimated magnitude spectrum to consist of a succession of randomly spaced spectral peaks [5]. This leads to an annoying residual noise, also called *musical noise* due to their narrow band spectrum and presence of tone-like characteristics. This noise although very different

from the original noise, can be very disturbing. A poorly designed spectral subtraction, can sometime results in a signal that is of a lower perceived quality and lower information content, than the original noisy signal. To eliminate the problem of musical noise and enhance spectral subtraction performance some modifications were introduced.

5. Modifications of spectral subtraction

5.1 Spectral subtraction using scaling factor and spectral floor

The first spectral subtraction method proposed by Boll [2] consists of implementation of the following relationship:

$$\left|\hat{X}(\omega)\right|^2 = \begin{cases} \left|Y(\omega)\right|^2 - \left|\hat{D}(\omega)\right|^2, \\ \quad if \ \ \left|Y(\omega)\right|^2 > \left|\hat{D}(\omega)\right|^2 \\ 0, \ \ otherwise \end{cases} \quad (6)$$

As it was discussed above, though the noise is reduced by this method, there is still considerable broadband noise (musical noise) remaining in the processed speech. To eliminate this problem the method proposed in [5] introduces two additional parameters to basic spectral subtraction algorithm. There are scaling factor α, and spectral floor β. Since the residual noise spectrum consists of peaks and valleys with random occurrences, spectral subtraction using scaling factor and spectral floor tries to reduce the spectral excursions for improving speech quality. This proposed technique can be expressed as:

$$\left|\hat{X}(\omega)\right|^2 = \begin{cases} \left|Y(\omega)\right|^2 - \alpha \cdot \left|\hat{D}(\omega)\right|^2, \\ \quad if \ \ \left|Y(\omega)\right|^2 - \alpha \cdot \left|\hat{D}(\omega)\right|^2 > \beta \cdot \left|\hat{D}(\omega)\right|^2 \\ \beta \cdot \left|\hat{D}(\omega)\right|^2, \ \ otherwise \end{cases} \quad (7)$$

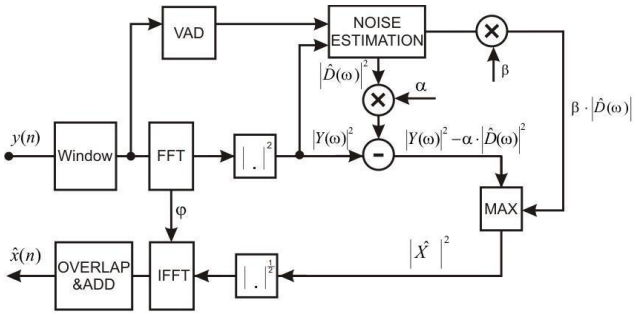

Fig. 3. Block diagram of modified spectral subtraction.

Where $\alpha \geq 0$, and $\beta << 1$. The harshness of the subtraction can be varied by applying a scaling factor α. The values of scaling factor α higher than 1 result in high SNR level of denoised signal, but too high values may cause distortion in perceived speech quality. Therefore, the value of α has to be chosen carefully in order to prevent both the musical noise and too much signal distortion. The introduction of spectral floor prevents the spectral components of the enhanced speech spectrum to descend below the lower bound $\beta \cdot \left|D(\omega)\right|^2$, thereby "filling-in" the deep valleys surrounding narrow peaks (from the enhanced spectrum). Reducing the spectral excursions of noise peaks (as compared to when the negative components are set to zero) reduces the amount of musical noise.

The performance of this type of SS algorithm is limited in the usage of stationary optimized parameters, which are difficult to choose for all speech and noise situations. It is difficult to suppress noise without decreasing intelligibility and without speech distortion, especially for very low signal-to-noise ratios.

5.2 Wiener filtration

It is convenient to consider the spectral subtraction as a filter, by manipulating (4) such that, it can be expressed as the product of noisy speech signal spectrum and the frequency response of a spectral subtraction filter as:

$$\left|\hat{S}(\omega)\right|^a = \left|Y(\omega)\right|^a - \left|\hat{D}(\omega)\right|^a = H(\omega) \cdot \left|Y(\omega)\right|^a \quad (8)$$

$$H(\omega) = 1 - \frac{\left|\hat{D}(\omega)\right|^a}{\left|Y(\omega)\right|^a} = \frac{\left|Y(\omega)\right|^a - \left|\hat{D}(\omega)\right|^a}{\left|Y(\omega)\right|^a} \quad (9)$$

The spectral subtraction filter is a zero phase filter, with its frequency response $H(\omega)$, is in the range of $0 < H(\omega) < 1$. The filter acts as a SNR-dependent attenuator. The attenuation in each frequency increases with the decreasing SNR, and vice-versa.

A transfer function of the Wiener filter [6], $H(\omega)_{wiener}$, is expressed in terms of the power spectrum of clean speech $P_s(\omega)$ and the power spectrum of noise $P_d(\omega)$ as in (9). But power spectrum of clean speech is not known, the power spectrum of the noisy speech $P_y(\omega)$ signal is used instead as:

$$H_{wiener}(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} = \frac{P_y(\omega) - P_d(\omega)}{P_y(\omega)} \quad (10)$$

Wiener filter cannot be applied directly to estimate the clean speech since speech cannot be assumed to be stationary. Therefore, an adaptive Wiener filter implementation can be used to approximate the above filter (10) as:

$$H_{wiener}(\omega) = \frac{E[|Y(\omega)|^2] - E[|D(\omega)|^2]}{E[|D(\omega)|^2]} \quad (11)$$

$$\left|\hat{S}(\omega)\right|^2 = H_{wiener}(\omega) \cdot |Y(\omega)|^2 \quad (12)$$

Comparing $H(\omega)$ and $H(\omega)_{wiener}$ from (9) and (11), it can be seen that the Wiener filter is based on the *ensemble average spectra* of the signal and noise, whereas the spectral subtraction filter (with a=2) uses the instantaneous spectra for noise signal and the running average (*time-averaged spectra*) of the noise. In Wiener filter theory the averaging operations are taken across the ensemble of different realization of the signal and noise processes. In spectral subtraction we have access only to single realization of the process.

Using of power spectrum of noisy speech, instead of that of clean speech for calculating the transfer function degrades Wiener filter accuracy. To solve this problem, an iterative algorithm is used [6]. In the algorithm the output signal of the Wiener filter is utilized to design a more accurate Wiener filter. Thus by iterating this process, we can design a high accurate Wiener filter. The input signal of the iterative Wiener filter is not renewed at each iteration. This means that only the filter is renewed.

5.3 Iterative spectral subtraction

To consider the musical noise problem common to conventional spectral subtraction method, an iterative spectral subtraction method was proposed in [7]. The iterative method is motivated by iterative Wiener filtering, where filtering output signal is used to design a higher performance Wiener filter. In iterative spectral subtraction the filtering output signal is used not only for designing the filter but also as the input signal of the next iteration process. Specifically for spectral subtraction, after the first spectral subtraction process, the type of additive noise is changed to that of musical noise. Then the noise signal is estimated from unvoiced segment parts. And, a new spectral subtraction filter is designed by using the new estimated noise (musical noise) and the new noisy speech (inc1uding the musical noise), which is the output signal by the first spectral subtraction. By the designed filter, an enhanced output signal can be obtained from the input signal. At every iteration musical noise is estimated in different frames, because the musical noise is not stationary in short time frames analysis. When we do such noise estimation, the spectral subtraction filter is always designed so as to reduce the musical noise remained in the previous spectral subtraction process. Therefore, the musical noise can be reduced significantly by performing the iterative spectral subtraction as shown.

5.4 Spectral subtraction based on perceptual properties

The choice of the subtraction parameters α, β and *a* is a main challenge in subtractive type speech enhancement algorithms. To track changes in background noise it is necessary to subtraction parameters to be adaptive. Good results are obtained, when the adaptation of subtractive parameters in time and frequency domain based on masking properties. Masking consists in the fact, that the human auditory system does not distinguish two signals when the signals are close to each other (in the time or frequency domain). In [8] the noise masking threshold $T(\omega)$ is used for adjusting spectral subtraction parameters α and β on a per frame and per frequency basis. The noise masking threshold is obtained through modeling the frequency selectivity of the human ear and its masking property. The different calculation steps are summarized in [8].

Therefore, the adaptation of subtractive parameters is based on the consideration, that if the masking threshold is high, residual noise will be masked and consequently be inaudible. Therefore, when the threshold is high, the subtraction parameters are kept minimal, thereby reducing speech distortion. When the masking threshold is low, the residual noise is not masked and the subtraction parameters are maximized. The following relations perform the adaptation of the subtraction parameters:

$$\begin{aligned} \alpha(\omega) &= F_\alpha[\alpha_{\min}, \alpha_{\max}, T(\omega)] \\ \beta(\omega) &= F_\beta[\beta_{\min}, \beta_{\max}, T(\omega)] \end{aligned} \quad (13)$$

where α_{min}, β_{min} and α_{max}, β_{max} are the minimal and maximal values of scaling factor and spectral floor respectively. F_α and F_β are the functions for a maximum reduction of residual noise: $F_\alpha=\alpha_{max}$ when $T(\omega)=T(\omega)_{min}$ and $F_\alpha=\alpha_{min}$ when $T(\omega)=T(\omega)_{max}$, where $T(\omega)_{min}$ and $T(\omega)_{max}$ are the minimal and maximal values, respectively, of the updated masking threshold. The values F_α between these two extreme limits are obtained by the interpolation of values $T(\omega)$. By similar considerations we obtain the values F_β. The following values were experimentally obtained to provide a good tradeoff for a human listener: $\alpha_{min}=1$ and $\alpha_{max}=6$; $\beta_{min}=0$ and $\beta_{max}=0.02$; exponent is

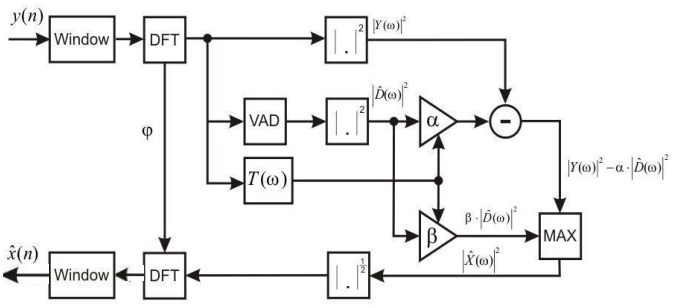

Fig. 4. Block diagram of a spectral subtraction with perceptual weighting

,constant a=2. Principle of SS with perceptual weighting is shown on Fig. 4.

6. Experiments and discussion

To compare speech enhancement techniques some experiments were made in Matlab environment. For experiments we have used speech signals from SpeechDat database [9] constituted by sentences pronounced in Czech language by male and female speakers. Sentesces were corrupted by two types of additive noise (AWGN and car noise) to obtain noisy speech with different values of the signal to noise ratio (SNR_{input}=15, 10, 5 and 0dB). The amount of noise reduction is generally measured with the SNR improvement, given by the difference between input and output segmental SNR. The obtained values of SNR improvement for two types of noise are given in Fig. 5. The best noise reduction is obtained in case of white Gaussian noise (AWGN), while for car noise this improvement decreases. For both types of noise, the SS with perceptual weighting and iterative SS achieve result in significant improvement over conventional SS. Modified SS and Wiener filtering outperform conventional SS on 1-2 dB. The greatest difference in algorithms performance can be observed in case of input signal at 0dB SNR level.

The main drawback of the SNR is the fact that it has a very poor correlation with subjective quality assessment results. SNR of enhanced speech is not sufficient objective indicator of speech quality. Structure of residual noise and speech distortion can be seen on spectrograms of denoised speech. Fig.6 represents spectrograms of speech enhanced by above described algorithms (conventional spectral subtraction (CSS), modified spectral subtraction (MSS) with scaling factor and spectral floor, Wiener filtration (WF), Iterative spectral subtraction (ISS) and spectral subtraction with perceptual weighting (SSPW)). As it shown on Fig. 6 conventional SS as well as modified SS contain audible residual noise, which can be annoying for listener. Wiener filtering results in a smaller amount of residual noise, but this noise has musical structure and speech regions, especially fricative consonants, are also attenuated. This type of SS can result in speech distortion.

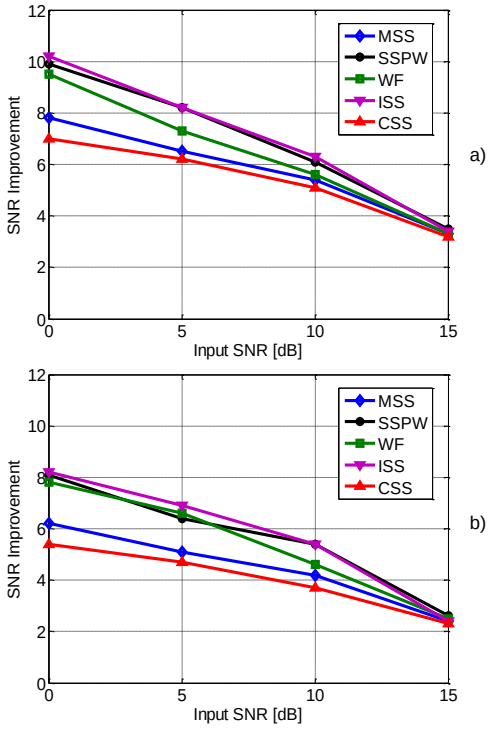

Fig. 5. SNR improvement of noise reduction algorithms for (a) AWGN noise, (b) Car noise.

The best results were obtained with SS algorithm with perceptual weighting. In case of this type of SS small amount of residual noise is leaved, but this noise has a perceptually white quality and distortion remains acceptable.

7. Conslusion

In this paper, some subtractive-type methods for acoustic noise reducing are introduced. In particular, methods based on short time Fourier transforms are examined. The limitations of spectral subtraction are briefly discussed. The artifacts introduced by SS methods are described, and how the conventional SS method is modified to counter these artifacts. From the SNR improvement point of view iterative SS and SS with perceptual weighting show the best noise reduction results from the other methods. Conventional SS, iterative SS and Wiener filtration algorithms results in audible residual noise, which can cause decreasing of speech intelligibility. The most progressive method of noise reduction is a SS with perceptual weighting based on masking properties of auditory model. This speech enhancement method takes advantage of how people perceive the frequencies instead of just working with SNR. It results in appropriate residual noise attenuating and acceptable degree of speech distortion.

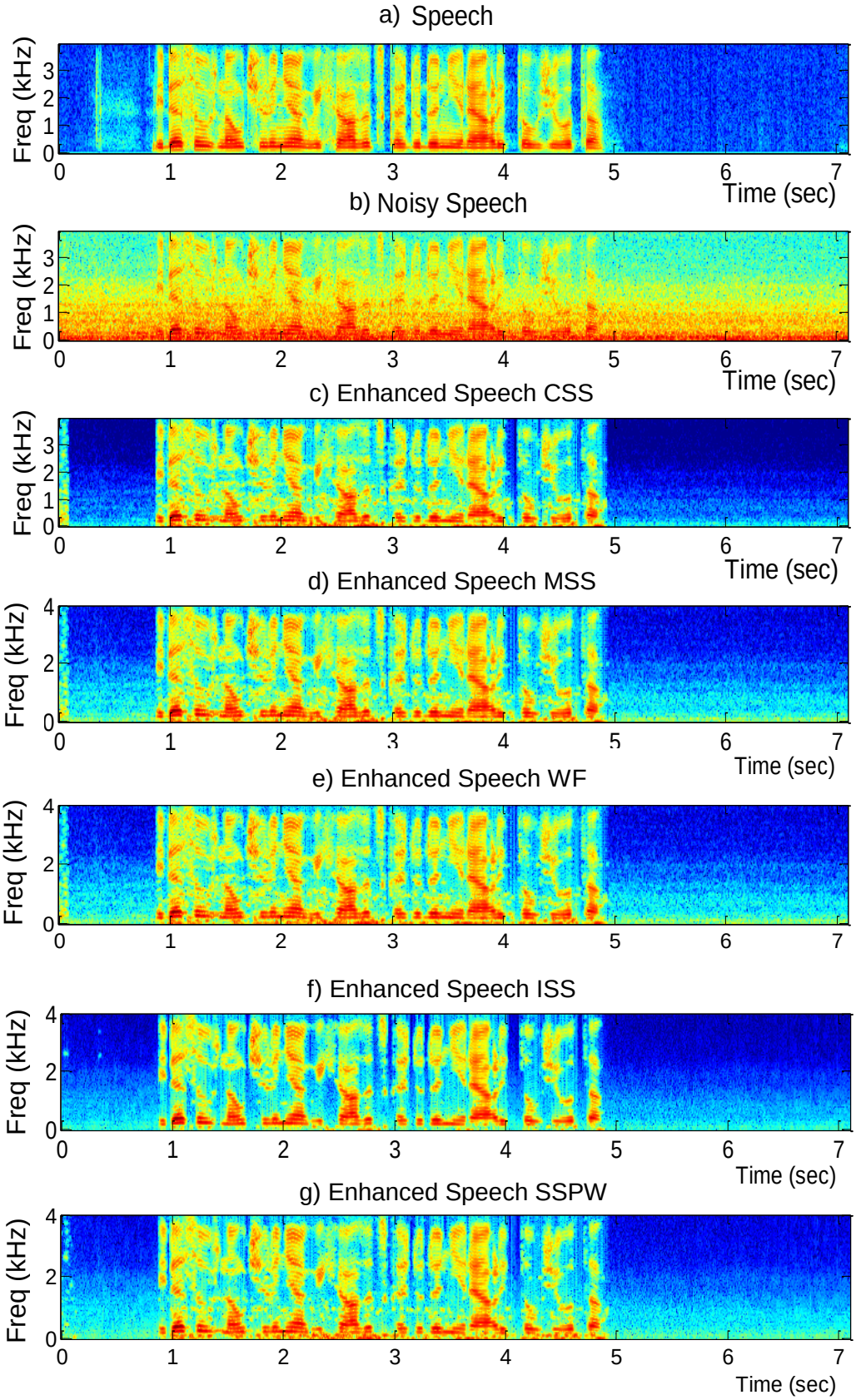

Fig. 6. Speech spectrograms. (a) Clean speech, (b) Noisy speech in the case of additive car noise (SNR = 0 dB), (c) – (g) Speech enhanced by noise reduction algorithms

Acknowledgements

Research described in the paper was supervised by Prof. Ing. B. Simak, CSc., FEL CTU in Prague and supported by Czech Technical University grant SGS No. OHK3-108/10 and the Ministry of Education, Youth and Sports of Czech Republic by the research program MSM 6840770014.

References

[1] Mourad T., "Simulation and Comparison of Noise Cancellation techniques in Speech Processing" Research Journal of Applied Science, vol 2, pp 119-123, Medwell Journals, 2007.

[2] Boll S., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. Acoust. Speech, Signal Processing , vol ASSP-27, pp 113-120, 1979.

[3] Sakhnov K. Verteletskaya E. "Dynamical Energy-Based Speech/Silence Detector for Speech Enhancement Applications" Proceedings of the World Congress on Engineering WCE 2009, Vol I, pp 801-806, 2009.

[4] Faneuff J.J, Brown D. R. "Noise Reduction and Increased VAD Accuracy Using Spectral Subtraction", Processing of the Global Signal Processing Exposition and International Signal Processing Conference (ISPC' 03). Dallas, Texas. April 2003.

[5] Berouti M., Schwartz R. and Makhoul J., "Enhancement of speech corrupted by acoustic noise", Proc. IEEE ICASSP, pp. 208-211, Washington DC, 1979.

[6] Lim, J.S. and Oppenheim, A.V., "Enhancement and bandwidth compression of noisy speech", Proc. IEEE, Vol. 67, No.12, pp. 1586-1604, 1979.

[7] Ogata, S.; Shimamura, T., "Reinforced spectral subtraction method to enhance speech signal" Proceedings of IEEE International Conference on Electrical and Electronic Technology, vol. 1, pp 242 – 245, 2001.

[8] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech and Audio Processing, vol. 7, pp. 126-137, 1999.

[9] Pollák, P.: Tvorba databází řečových signálů pro účely rozpoznávání a zvýrazňování. [Docentská habilitační práce]. Praha: ČVUT v Praze, FEL,104 s, 2002.

About Authors ...

Ekaterina VERTELETSKAYA was born in Uzbekistan. She obtained the MSc. degree in Telecommunication and Radio engineering from Czech Technical University, Prague, in 2008. Currently she is a Ph.D. student at the Department of Telecommunication Engineering of the CTU in Prague. Her research activities are in the area of speech signal processing, focused on noise reduction.

prof. Ing. Boris ŠIMÁK, CSc. is actively involved in research of digital signal processing in the area of speech and image processing. He is the technical director of R&D centre for mobile communication at CTU in Prague and the member of executive board of Sitronics Centre at CTU in Prague. Since 2007, he is the dean of Faculty of Electrical Engineering, CTU in Prague. He participated on several national (GAČR, FRVŠ, MPO) and international projects (TEMPUS, LEONARDO).